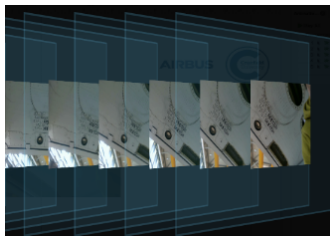


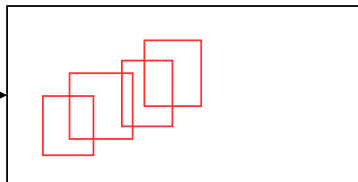
Input Video Frames
Sequence $S_{t-1+n:t}$



Object Detection Model

Observed Bounding Boxes

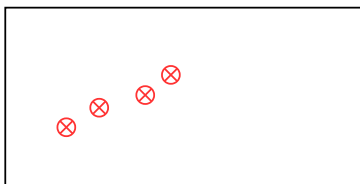
$(x_{\text{center}}, y_{\text{center}}, w, h)$



Data Preprocessing

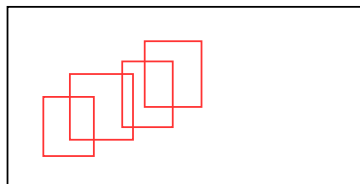
Observed Spatial Dynamics

$(x_{\text{center}}, y_{\text{center}}, v_x, v_y, a_x, a_y)$



Observed Dimensional Attributes

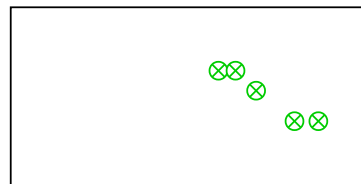
$(w, h, \Delta w, \Delta h)$



Sequence Model

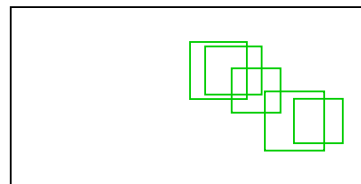
Predicted Spatial Dynamics

$(x'_{\text{center}}, y'_{\text{center}}, v'_x, v'_y, a'_x, a'_y)$

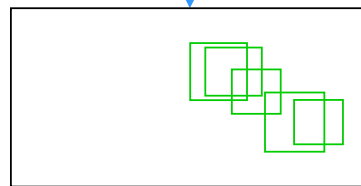


Predicted Dimensional Attributes

$(w', h', \Delta w', \Delta h')$



Data Postprocessing



Output Bounding Boxes Sequence
 $S_{t+1:t+m}$

$(x'_{\text{center}}, y'_{\text{center}}, w', h')$